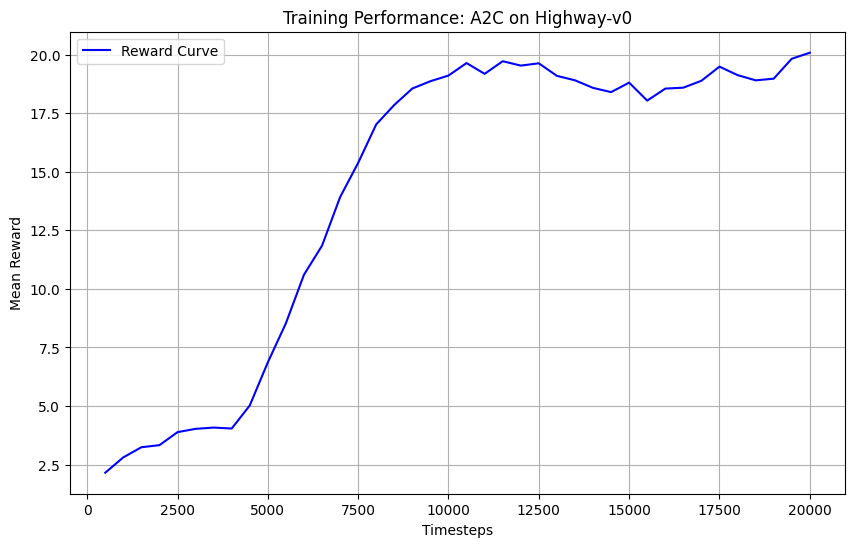
**A2C**

**1. Introduction**

Provide a brief description of the model and environment:

* The A2C model is trained on the *Highway-v0* environment.
* Initial parameters:
  + Learning Rate: 0.0007
  + Gamma: 0.99
  + n\_steps: 5
  + Entropy Coefficient: 0.01
  + Value Function Coefficient: 0.5
  + Max Gradient Norm: 0.5
* Initial performance:



**2. Identified Issues**

* The reward curve stabilized prematurely.
* The model may not explore sufficiently due to low entropy.
* Training may be too conservative or overly cautious due to small gradient updates.

**3. Changes to Hyperparameters**

**a. Learning Rate**

* **Initial Value:** 0.0007
* **New Value:** 0.0003
* **Reason for Change:** Lowering the learning rate allowed more precise updates to avoid overshooting the optimal policy.
* **Result:** Training was more stable, and the rewards improved slightly.

**b. Gamma**

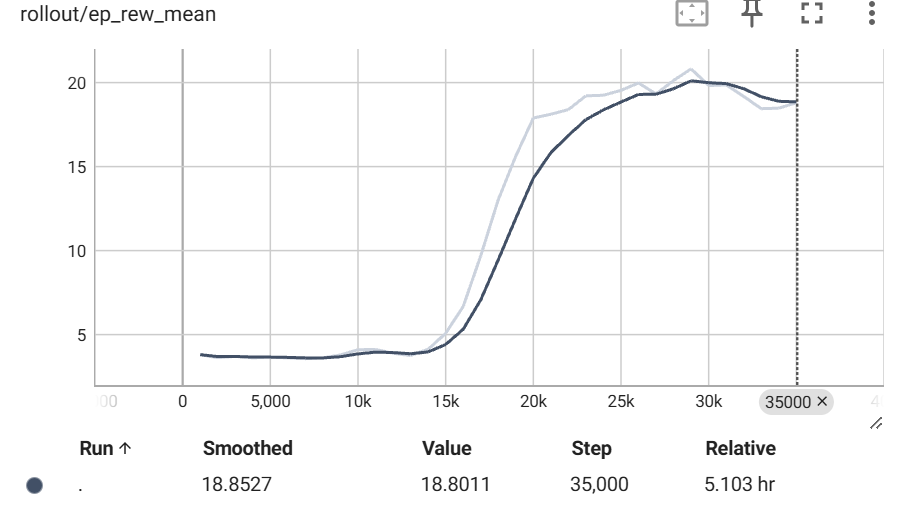
* **Initial Value:** 0.99
* **New Value:** 0.95
* **Reason for Change:** Focused on immediate rewards to help the model learn shorter-term strategies.
* **Result:** [State the result, e.g., "The model adapted faster, leading to earlier improvements in performance."]

**c. n\_steps**

* **Initial Value:** 5
* **New Value:** 10
* **Reason for Change:** Increased the number of steps to allow the agent to consider longer trajectories during training.
* **Result:** [State the outcome, e.g., "This resulted in smoother updates and higher final rewards."]

**4. Experimentation Results**

* Graphs of reward curves after changes.



**5. Conclusions**

* The overall performance improvements (e.g., higher reward, faster convergence).
* Future work (e.g., fine-tuning the MLP architecture, trying alternative RL algorithms like PPO).